A Data-Science-as-a-Service Model

Matthias Pohl, Sascha Bosse and Klaus Turowski
Magdeburg Research and Competence Cluster,
Faculty of Computer Science, University of Magdeburg, Magdeburg, Germany.

Keywords: Data-Science-as-a-Service, Cloud Computing, Service Model, Data Analytics, Data Science.

Abstract: The keen interest in data analytics as well as the highly complex and time-consuming implementation lead to an increasing demand for services of this kind. Several approaches claim to provide data analytics functions as a service, however they do not process data analysis at all and provide only an infrastructure, a platform or a software service. This paper presents a Data-Science-as-a-Service model that covers all of the related tasks in data analytics and, in contrast to former technical considerations, takes a problem-centric and technology-independent approach. The described model enables customers to categorize terms in data analytics environments.

1 INTRODUCTION

An increasing keen interest in data science and data analytics exists as reported by a trend analysis of the last 5 years (Google Trends). The data-intensive world allures with generating revenue from analyzing information and data that are simply and quickly available. Different approaches for knowledge discovery (KDD) (Fayyad et al., 1996) or business-related data mining (CRISP-DM) (Shearer, 2000) are in use. However, proceeding is highly complex, time-consuming and needs expertise in different disciplines, either computer sciences, mathematics or a context-related specialization. The motivation within a company could arise from preventing downtime of machines, getting insights about customer relationships or optimizing business processes. If the required expertise for data analysis cannot be provided internally, the tasks can be forwarded to external consulting services. By using such services it is possible to compensate the lack of expertise, but it is very cost-intensive and still extremely time-consuming. The paradigm of cloud computing (Mell and Grance, 2011) establishes service concepts that seem to be able to solve the remaining issues. Among concepts like Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS) or Software-as-a-Service (SaaS) that revolutionize computing on several layers, service models like Analytics-as-a-Service, Data-Analysis-as-a-Service, Business-Analytics-as-a-Service or Big-Data-as-Service were conceptualized. Whether these approaches are suitable recommendations, decision or application services for non-expert customers or just analogies to standard concepts is not clarified. Furthermore, a customer is confronted to choose within a set of apparently unclear buzzwords like data science, data mining, big data analytics, etc.

Therefore we make a predefinition for the usage of these terms and will diffuse a definition through argumentation. The whole process that contains data provision, data preparation, data analysis and data visualization is called data science in this paper. Data analytics, business analytics and big data analytics are often used synomously for data science, however they could differ in context of use. Data mining is a key term in most of related works and is used in varient different ways. We will take this term as a byword for data analysis.

This paper has twofold objectives. Firstly, it will provide a full-service model that will be extracted from existing approaches and will address data analytics services. Secondly, it will discuss the arising service offers and data science process steps. The subsumption of service models simplifies the assessment of IT service offers for companies that plan to get insight from their data. From a scientific view, a guideline is drawn for a future usage of terms in data analytics environment and a classification of past work is a point of interest. The structure of the paper is as follows. A knowledge base about related work is presented in the second section. The review is built up on a cross-reference search for data analytics services and data mining process steps (Webster et al.,

Pohl, M., Bosse, S. and Turowski, K.
A Data-Science-as-a-Service Model.
DOI: 10.5220/0006703104320439
In *Proceedings of the 8th International Conference on Cloud Computing and Services Science (CLOSER 2018)*, pages 432-439
ISBN: 978-989-758-295-0
Copyright © 2019 by SCITEPRESS – Science and Technology Publications, Lda. All rights reserved

2002). The third section shows a Data-Science-as-a-Service model that combines the different steps in data mining as well as a cloud computing service model. Section four demonstrates examples of existing services that are offered by known IT service providers. A discussion that argues aspects and challenges from a technical and a scientific view as well as concludes the paper.

2 RELATED WORKS

There are numerous works that address data mining, data analytics, data science or similar. First of all, we want to have a look on frameworks that are related to data mining, data analytics and data science. In case of cloud computing, different services that provide data analytics are already developed. Next to data analytics services we will consider data science process steps and focus on it in further search. In order to retrieve relevant works, we took search engine (Science Direct, Scopus) and knowledge databases (DBLP, IEEE, ACM) into account.

2.1 Data Science Frameworks

The data science process shall end up with knowledge. The key step in this nontrivial process is called data mining (Fayyad et al., 1996). Therefore data has to be processed from selected sources and transformed into a proper format. CRISP-DM (Shearer, 2000) describes a process cycle that is similar to Knowledge Discovery in Databases (Fayyad et al., 1996), but points out the business understanding that is necessary for initiating because data mining goals will be defined on that basis. Respecting both frameworks it is conspicuous that data mining is a name of a process step in KDD and entitles the whole reference model CRISP-DM. In (Kurgan and Musilek, 2006), a review of data mining frameworks is given. There exist various derivates of the mentioned frameworks with relation to a specific field of application. In (Sun et al., 2017), the authors add data analytics that also contains data mining as a significant part to the discussion and observe it as an aggregation of data analysis, data mining, data warehouse, statistic modeling and data visualization. Furthermore the authors aim at identifying an ontological separation between data analytics, business analytics, knowledge analytics and big data analytics. In (Nalchigar and Yu, 2017), a conceptual modeling framework for business analytics is proposed and provides catalogues for business questions, analytic algorithms and data preparation. In this manner, a transmission between business requirements and algorithmic implementation is established. With the evolution of a data-intensive world new technologies are needed for handling rapidly growing, variously shaped mass of data (Laney, 2001). In (Maltby, 2011), the author reviews big data with relation to analytic techniques and mentions that data mining combines statistics and machine learning with database management. Machine learning is not a new thing (Michalski et al., 1983) and focuses on ”automatically learn to recognize complex patterns and make intelligent decisions based on data” (Maltby, 2011). Data analysis, machine learning and data visualization are marked as the core disciplines of data science that is a new paradigm in the field of big data (Concolato and Chen, 2017). However, there is no simple and unified data science framework (Marvasti et al., 2015). Such frameworks are often deep into specific context-related data and implemented in an application environment (Brough et al., 2017; Loetsch and Ultsch, 2016). In (Cleveland, 2001), data analysis is described as an enlargement of statistics combined with data computing and is called data science. Obviously, there exists discordance about the terms data science, (big) data analytics, data mining, etc. However, it occurs that all definitions come up with a similar structure. Data has to be selected, prepared and analyzed to show up new information. However, the needed expertise and the lack of uniformed processes lead to the fact that data analytics is more an art than science (Zorrilla and García-Saiz, 2013).

2.2 Data Science Services

The thought of automation without requiring human interaction, the capability solving small and big data problems, the availability as well as flexibility in computing power and storage characterize the essentials of cloud computing (Mell and Grance, 2011). Different concepts that offering machines (IaaS), operating systems (PaaS) and applications (SaaS) are already defined and discussed (Dillon et al., 2010). There exist approaches for hosting data analytics environments in a cloud. In (Xu et al., 2015), the authors aim at making real-time data analytics available as a service and deal with the challenges of creating service interfaces, wrapping existing big data frameworks and real-time processing of data. In (Zorrilla and García-Saiz, 2013), an approach is motivated with some exemplary services (e.g. GoodData, IBM Smart Analytics) and it is concluded that none of them wraps the whole analytic process that is described in KDD or CRISP-DM. The proposed framework contains services for the steps of data preparation, data analysis and data visualization and is layer structured from an en-

terprise perspective. The concept is implemented as a SaaS, though its automation is not sufficiently discussed. In (Ribeiro et al., 2015), the authors provide a Machine-Learning-as-a-Service approach and discuss the low flexibility of implementing new algorithms on existing platforms like PredictionIO or OpenCPU. In (Xinhua et al., 2013), Big-Data-as-a-Service is defined with getting insight from big data, which characterizes this as big data analytics. In (Ardagna et al., 2016), the authors describe the Big-Data-Analytics-as-a-Service paradigm as suitable of companies do not have enough data scientists. Along to service requirements (data preparation, data analysis, data visualization), various challenges (data quality, diversity, security, privacy) that come up with handling vast data are examined by the authors. In (Grossman et al., 2016) the variations of data science services (infrastructure, platform, software, support) and a collocation of these is mentioned. The authors also describe requirements like API-based access, data portability, data peering and pay-for-compute, however automation in processing is not addressed. In (Medvedev et al., 2017), an approach for setting up data mining processes in a cloud environment is proposed and allows a user to model a data mining process within the steps of data uploading, data pre-processing, data mining and presentation. Several Analytics-as-a-Service providers are compared in (Naous et al., 2017) with respect to core features of data sources, data processing & preparation, analysis, visualizations and even platform and infrastructure propositions. In conclusion all provided services have a self-service character.

2.3 Data Science Process Steps Services

After terms and existing data science services have been discussed, we want to focus on the steps that represent the parts of the data science process. Companies are often interested in analyzing internal and external business-related data. Within the scope of digitalization, Internet-of-Things (IoT) applications are in widespread use. In conjunction with that, a lot of data services are formed that could have their origin in data crawling concepts. In (Haase et al., 2011), an information workbench for linked data applications is provided, and a structured crawling system to gather linked data over the web with standards like RDF is proposed in (Isele et al., 2010). However such services are known and described as Data-as-a-service (Delen and Demirkan, 2013). In (Terzo et al., 2013), Data-as-a-Service is interpreted as a storing and processing service and the authors proposed a layer architecture of an IaaS. In (Seibold and Kemper, 2012), different types of Database-as-a-Service (shared machine, shared processes, shared tables) are described and can appear as SaaS, PaaS and IaaS, which depends on the complexity of the delivered systems. In (Curino et al., 2011), the authors point out some properties (efficient multi-tenancy, elastic scalability, and database privacy) that have to be realized which is confirmed in (Agrawal et al., 2009). With respect to data storing, data integration is similarly important. In (Riedemann and Timm, 2003), the authors draw the ”vision to achieve automated just-in-time integration”, and in (Bergamaschi et al., 2008) the idea comes up that ”data integration systems is on producing a comprehensive global schema successfully integrating data from heterogeneous data sources”. In (Cohen and Richman, 2002), it is mentioned that data matching and clustering algorithms can create a solution for these problems. At this point an intersection with data preparation is obvious. ”Data preparation is an important and critical step for complex data analysis” (Yu et al., 2006). A handbook for data preparation is provided in (Pyle, 1999) and describes the access of data, data discovery and data modeling. It also focuses on data mining and determines that ”the process of data science is smooth and backward adjustment is possible”. In (Yu et al., 2006), the authors mention problems in data preparation like incomplete data, noisy data, inconsistent data, selecting relevant data, reducing data and resolving data conflicts, however, they notices some solution approaches. In (Narman et al., 2009), a model-based method for detecting data accuracy problems is proposed. The straightened data has to be organized, so data modeling is a further sub-process. In (Duggan and Yao, 2015), the authors describe ”an approach [for] automating most of this work, building data models from specifications of a data collection system”.In (Song et al., 2015), the authors address problems like combinatorial complexity, scattered modeling rules, semantic mismatch, inexperience of novice designers, incomplete knowledge of designers and multiple solutions in data modeling and give some solution techniques that are categorized in linguistics based (e.g. NLP), pattern-based, case-based, ontology-based and multi-techniques-based. At this point, one can see a relation to machine learning techniques that can also be used for outlier detection (Hawkins et al., 2002; Pruengkarn et al., 2016) or data matching (Rong et al., 2012). The overall modeling concept could be named as data warehousing. A transformation approach from operational schemes to data warehouse is presented in (Dori et al., 2008). Such a data structure changes with time and arising amount of data, so a framework for real time data warehousing is suitable

(Farooq and Sarwar, 2010). With the usage of machine learning techniques, one can see a connection to data analysis tasks. The main aim is to automate the task of selecting algorithms that analyze the data. In (García-Saiz and Zorrilla, 2017), a framework is presented that includes a meta-learning approach. With model-based, data contextual, information theoretical, complexity and statistical meta features a system learns the way of appropriate algorithm selection for a common problem. In (Luo, 2016), the author provides a literature review on automatic algorithm selection for machine learning and categorizes different approaches. It also exposes that the parameter and feature selection is an important part. In (Langley et al., 1994), an approach with a heuristic search while in (Hall, 2000) a correlation-based feature selection is described. A different approach is introduced in (Espinosa et al., 2013), in fact a taxonomy that provides a recommendation about using data mining methods for non-expert data miners. After these steps, a suitable visualization or presentation of the results is necessary. In (Matsushita et al., 2004), the authors focus on an automated visualization of user required information via processing data frames. In (Andrienko and Andrienko, 1999), a data characterization scheme for automated data visualization is provided. The cycle between data analysis and data visualization is picked up in (Wagner, 2015) where a process of reconfiguring data analysis with ideas coming from visualizations is introduced. The essential challenges are concluded in (Xu et al., 2015). The creation of service interfaces that are suitable for integrating and processing distinct data sources in combination with analyzing frameworks and that are also utilizable for different applications is paired with the ability of real-time responding.

3 DATA-SCIENCE-AS-A-SERVICE MODEL

Considering prior works in the field of data science and correlated disciplines, a Data-Science-as-a-Service model will be deduced in this section. The majority of the existing approaches provide IaaS, PaaS or SaaS models that enable users to conduct data science. However, for serving data science there has to be a service level above the known cloud computing models (Mell and Grance, 2011) that covers the whole data science process. The earlier frameworks and theoretical data mining approaches overlap in the core steps of accessing data, arranging data, analyzing data and presenting results. The automatization of the process steps leads to separated services that can be used as standalone ones. Therefore, we will describe some entry and exit points of the sub-services to show the feasibility of service separation. An entry point is referred to a condition for initiating a sub-process. An exit point is defined as a point at which results are available.

3.1 Data-as-a-Service

There have to exist data (Fayyad et al., 1996; Shearer, 2000) for a data science process. Either data is uploaded or collected by a user (Naous et al., 2017; Medvedev et al., 2017) or integrated from data services (Delen and Demirkan, 2013) in an adequate storage system. The aim of this step called Data-as-a-Service is providing a base of data that ideally cover all related information. A first entry point of this service model is a chunk of data or an information request that could be represented by keywords. On basis of the extracted meta data or keywords related data could be provided via data services. Hence, an exit point is the provision of required data or an index-linked data pool that is, for instance, stored in a traditional database or a distributed file system. It is shown that data services can be connected even with other data sources and supplied on appropriate infrastructure (Vu et al., 2012; Isele et al., 2010; Bergamaschi et al., 2008).

3.2 Data-Preparation-as-a-Service

The process step that follows after data provision is data preparation. It is termed as data selection, data preprocessing, data transformation (Fayyad et al., 1996), directly data preparation (Ardagna et al., 2016; Naous et al., 2017; Shearer, 2000), data modeling or generally data warehousing. However, all of these terms are covered by Data-Preparation-as-a-Service. Next to the previous service step a user-provided data pool could also be an entry point for this service, however, one has to consider that at least infrastructure (IaaS) is needed. An exit point is a fully organized set of data that is suitable for data analysis. In case of unstructured or semi-structured data the data preparation service step is necessary before starting the data analysis service. Following (Duggan and Yao, 2015; Song et al., 2015; Yu et al., 2006; Narman et al., 2009) it is possible to arrange data automatically after cleansing.

3.3 Data-Analysis-as-a-Service

The key step in the majority of the related approaches is called data mining (Fayyad et al., 1996; Medvedev et al., 2017; Shearer, 2000; Zorrilla and García-Saiz,

2013; Sun et al., 2017). Data analysis is also used for indication (Xinhua et al., 2013; Naous et al., 2017; Sun et al., 2017). The term data analytics that we introduced as a synonym for data science is seldomly mentioned (Ardagna et al., 2016), just like (machine) learning and predicting (Ribeiro et al., 2015). However, all the approaches use the referred terms in sense of Data-Analysis-as-a-Service. The concept of data mining transposes the idea of creating data by predictive or prescriptive analysis. Nevertheless, in general it is a derivation of data analysis. In a standalone service consumption an entry point could be a well-structured data set that will be used as input data for analysis procedures and algorithms. An evaluated output (e.g. a data frame) is a possible exit point. The automatic selection of algorithms (Luo, 2016) and a self-learning application system (García-Saiz and Zorrilla, 2017) enable such a service.

3.4 Data-Visualization-as-a-Service

The last service step adresses data visualization which is a term that has been chosen by most researchers (Ardagna et al., 2016; Xinhua et al., 2013; Naous et al., 2017; Sun et al., 2017; Zorrilla and García-Saiz, 2013). However, an interpretation of results (Fayyad et al., 1996), a deployment (Shearer, 2000) or a structured output for a customers product (Xu et al., 2015) is mentioned. An entry point could be structured data that should only be visualized. However, a service for distributing analysis results is also conceivable that could be seen as the exit point. Approaches by (Matsushita et al., 2004) and (Andrienko and Andrienko, 1999) can facilitate a visualization service.

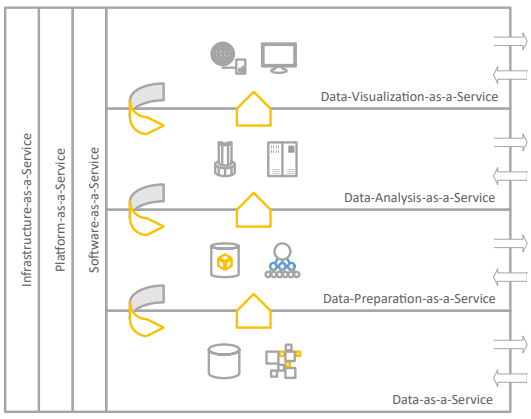

Figure 1: Data-Science-as-a-Service model.

3.5 Data-Science-as-a-Service

The Data-Science-as-a-Service model (Fig. 1) combines the explained services and realizes the whole data science process. Data-Science-as-a-Service is the ordered sequence of Data-as-a-Service, Data-Preparation-as-a-Service, Data-Analysis-as-a-Service and Data-Visualization-as-a-Service. The entry points of Data-Science-as-a-Service are equal to the ones of Data-as-a-Service. Exit points could be derived from Data-Visualization-as-Service. Considering the model a combination that does not contain all of the sub-processes is also conceivable. The entry and exit points of the sub-services enable the usage of sub-sequences. For instance, one could getting started with a information requirement at Data-as-a-Service and quit with a well-structured dataset at Data-Preparation-as-a-Service.

Furthermore, the idea of backward adjustment (Pyle, 1999; Wagner, 2015) is included. In case a service result would not fulfill the customer requirements a service step could be re-processed or a previous service step could be invoked. For instance, if the output result of the data analysis service is not suitable or applicable then the step could be repeated. At this point the service system gets the possibility to learn and can forward the demand of advancement if it is not successful. Otherwise the system rolls back to the underlying Data-Preparation-as-a-Service to rearrange the data (e.g. new metrics) to gain better results in the next step. This functionality is only useable in case of requesting more than one service.

The essential characteristics of cloud computing (Mell and Grance, 2011) are necessary for providing Data-Science-as-a-Service. Although the database services or data integration services are part of Data-as-a-Service these services have to be supported by infrastructure, platform and software to process all of the service steps. At this point we do not want to categorize the underlying services. Nevertheless, resource elasticity, service measurement and broad network access is required to offer Data-Science-as-a-Service.

Data analytics could be used as a similar expression for data science. Normally the initiation of a data science process is forced by a user or business requirements that expect some kind of knowledge or insight from data. Thus, the terms business analytics, Insight-as-a-Service or Knowledge-as-a-Service (Terzo et al., 2013) are used. In case of big data problems, there exist different approaches about so called Big-Data-Analytics-as-a-Service or Big-Data-as-a-Service. However, all of these concepts are covered by the presented Data-Science-as-a-Service concept.

4 EXAMPLE

In the previous section we presented a Data-Science-as-a-Service model. In (Naous et al., 2017), it is shown that all of the data science cloud computing offers (e.g. Google Cloud Platform, IBM Bluemix, SAP Cloud Platform, Amazon Quicksight or GoodData Platform) are some kind of self-service via software or platform. The GoodData and SAP Cloud platform cover the most service steps, from data provision up to visualization. However, if one has a look on the current product line of Google Services all data science process steps are covered (e.g. BigQuery, CloudStorage, DataFlow, TensorFlow, Prediction API, Charts and Firebase). Nevertheless none of the listed providers offers a completely data science service that proceeds (nearly) automatically.

Observing Google's image recognition tools we find a demonstrating example for Data-Science-as-a-Service. For instance, Google Goggles combines the whole sequence of the Data-Science-as-a-Service model. The starting point of the service is a taken picture of a user. Google provide databases with website links and images as Data-as-a-Service. The image of a user has to be transformed in a suitable format for further processing (Data-Preparation-as-a-Service). Google's machine learning algorithms will detect patterns of interest by means of, for instance, trained neural networks (Data-Analysis-as-Servce). Afterwards a presentation of the detected image areas and referred website links is given (Data-Presentation-as-a-Service). Furthermore, the service can learn from a rating of the result regarding if a user is satisfied or not. In the unsatisfied case enriching the databases or implementing new algorithms could be fitting solutions. New computation methods may assume reprepared input data. Although, new data definitely involve a repreparation. Hence, we observe the formerly mentioned backward adjustment.

5 DISCUSSION AND CONCLUSION

There is a many-faceted choice of products that range from IaaS over PaaS to SaaS. Providing a platform or a software that allows to conduct data science is not Data-Science-as-a-Service, however such an offer is not possible without. Data-Science-as-a-Service is a symbiosis of infrastructure, platform, software and the processing of data science tasks. This can be done automatically or semi-automatically, e.g. with options of user interactions for launching further service steps or re-processing. Otherwise there would not be an added value for a customer in comparison to common services. Considering (Vargo and Lusch, 2004), value only results from the beneficial application of data science services where the infrastructure is a transmitter. Furthermore, the essential characteristics of cloud computing like on-demand self-service, broad network access, resource pooling, rapid elasticity and measured service are given. The same refers to the deployment models. Considering the service measurement one could think about new business models that specify a typical pay-per-use model. Among a pay-per-compute, a pay-per-insight model is potentially conceivable. Measurements could also differ on service levels. Ultimately, it is an existing future task, because it drops the question how to measure insight in that case. We have shown that a data science service system is feasible with the connection of several service steps. However we describe an overall service model and not the orchestration of services in detail. The connection between the service steps has to be observed in future. There exist different frameworks (e.g. RDF) or markup languages (e.g. PMML) that can be used to determine utility. Challenges will arise to achieve smooth transitions, even though when a service step is skipped by a customer or the functionality of backward adjustment is applied. Therefore it is not even called service layers, because the usage as stand-alone services is also possible. "Information systems that provide such capabilities are often called business intelligence tools nowadays" (Delen and Demirkan, 2013). Indeed one can see parallels to business intelligence (BI). Comparable concepts are given in the theory of decision support systems (DSS) that could be included in future research. Especially BI tools focus on a technical layer structure for collecting business data, re-arrangement and reporting. However, there also exists an alternate definition that sees analytics as a subset of BI (Davenport et al., 2001). From a point of service a data/business understanding is not necessary. The customer is forced to input its information demand and to check the knowledge outcome on its suggestions. If an automated data science service system is able to govern requirements and in-depth knowledge one might call it Artificial Intelligence. Associating decision support decision model presentation could be included in Data-Presentation-as-a-Service in future. With the given service models that orientate towards the key steps of data mining it is possible to characterize terms that occurs in the analytics service field. One has only to decide if a term is related to a sub-process or the whole system. Provisioning

an ontology in the field of data analytics will be a prospective aim.

In this paper, a service model framework for data science was created. It enables business and scientific customers to classify offers of common data science services and to substantiate their expectations. The results are furthermore useful as a template for creating data science services.

REFERENCES

Agrawal, D., El Abbadi, A., Emekci, F., and Metwally, A. (2009). Database management as a service: Challenges and opportunities. In *Data Engineering, 2009. ICDE'09. IEEE 25th International Conference on*, pages 1709–1716. IEEE.

Andrienko, G. and Andrienko, N. (1999). Data characterization schema for intelligent support in visual data analysis. *Lecture notes in computer science*, pages 349–366.

Ardagna, C. A., Ceravolo, P., and Damiani, E. (2016). Big data analytics as-a-service: Issues and challenges. In *Big Data (Big Data), 2016 IEEE International Conference on*, pages 3638–3644. IEEE.

Bergamaschi, S., Po, L., and Sorrentino, S. (2008). Automatic annotation for mapping discovery in data integration systems. In *SEBD*, volume 2008, pages 334–341.

Brough, D. B., Wheeler, D., and Kalidindi, S. R. (2017). Materials knowledge systems in python—a data science framework for accelerated development of hierarchical materials. *Integrating Materials and Manufacturing Innovation*, 6(1):36–53.

Cleveland, W. S. (2001). Data science: an action plan for expanding the technical areas of the field of statistics. *International statistical review*, 69(1):21–26.

Cohen, W. W. and Richman, J. (2002). Learning to match and cluster large high-dimensional data sets for data integration. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 475–480. ACM.

Concolato, C. E. and Chen, L. M. (2017). Data science: A new paradigm in the age of big-data science and analytics. *New Mathematics and Natural Computation*, 13(02):119–143.

Curino, C., Jones, E. P., Popa, R. A., Malviya, N., Wu, E., Madden, S., Balakrishnan, H., and Zeldovich, N. (2011). Relational cloud: A database-as-a-service for the cloud. In *5th Biennial Conference on Innovative Data Systems Research*, pages 235–240. MIT.

Davenport, T. H., Harris, J. G., David, W., and Jacobson, A. L. (2001). Data to knowledge to results: building an analytic capability. *California Management Review*, 43(2):117–138.

Delen, D. and Demirkan, H. (2013). Data, information and analytics as services. *Decision Support Systems*, 1(55):359–363.

Dillon, T., Wu, C., and Chang, E. (2010). Cloud computing: issues and challenges. In *Advanced Information Networking and Applications (AINA), 2010 24th IEEE International Conference on*, pages 27–33. IEEE.

Dori, D., Feldman, R., and Sturm, A. (2008). From conceptual models to schemata: An object-process-based data warehouse construction method. *Information Systems*, 33(6):567–593.

Duggan, D. and Yao, J. (2015). Automated data modeling for health data collection. Stevens Institute Technical Report.

Espinosa, R., García-Saiz, D., Zorrilla, M., Zubcoff, J. J., and Mazón, J.-N. (2013). Enabling non-expert users to apply data mining for bridging the big data divide. In *International Symposium on Data-Driven Process Discovery and Analysis*, pages 65–86. Springer.

Farooq, F. and Sarwar, S. M. (2010). Real-time data warehousing for business intelligence. In *Proceedings of the 8th International Conference on Frontiers of Information Technology*, page 38. ACM.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., et al. (1996). Knowledge discovery and data mining: Towards a unifying framework. In *KDD*, volume 96, pages 82–88.

García-Saiz, D. and Zorrilla, M. (2017). A meta-learning based framework for building algorithm recommenders: An application for educational arena. *Journal of Intelligent & Fuzzy Systems*, 32(2):1449–1459.

Grossman, R. L., Heath, A., Murphy, M., Patterson, M., and Wells, W. (2016). A case for data commons: Toward data science as a service. *Computing in Science & Engineering*, 18(5):10–20.

Haase, P., Schmidt, M., and Schwarte, A. (2011). The information workbench as a self-service platform for linked data applications. In *Proceedings of the Second International Conference on Consuming Linked Data-Volume 782*, pages 119–124. CEUR-WS.org.

Hall, M. A. (2000). Correlation-based feature selection for discrete and numeric class machine learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 359–366. Morgan Kaufmann Publishers Inc.

Hawkins, S., He, H., Williams, G., and Baxter, R. (2002). Outlier detection using replicator neural networks. In *DaWaK*, volume 2454, pages 170–180. Springer.

Isele, R., Umbrich, J., Bizer, C., and Harth, A. (2010). Ldspider: An open-source crawling framework for the web of linked data. In *Proceedings of the 2010 International Conference on Posters & Demonstrations Track-Volume 658*, pages 29–32. CEUR-WS.org.

Kurgan, L. A. and Musilek, P. (2006). A survey of knowledge discovery and data mining process models. *The Knowledge Engineering Review*, 21(1):124.

Laney, D. (2001). 3d data management: Controlling data volume, velocity and variety. *META Group Research Note*, 6:70.

Langley, P. et al. (1994). Selection of relevant features in machine learning. In *Proceedings of the AAAI Fall symposium on relevance*, volume 184, pages 245–271.

Loetsch, J. and Ultsch, A. (2016). Process pharmacology: A pharmacological data science approach to drug development and therapy. *CPT: Pharmacometrics & Systems Pharmacology*, 5(4):192–200.

Luo, G. (2016). A review of automatic selection methods for machine learning algorithms and hyper-parameter values. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 5(1):18.

Maltby, D. (2011). Big data analytics. In *74th Annual Meeting of the Association for Information Science and Technology (ASIST)*, pages 1–6.

Marvasti, M. A., Poghosyan, A. V., Harutyunyan, A. N., and Grigoryan, N. M. (2015). Ranking and updating beliefs based on user feedback: Industrial use cases. In *2015 IEEE International Conference on Autonomic Computing*, pages 227–230.

Matsushita, M., Maeda, E., and Kato, T. (2004). An interactive visualization method of numerical data based on natural language requirements. *International journal of human-computer studies*, 60(4):469–488.

Medvedev, V., Kurasova, O., Bernatavičienė, J., Treigys, P., Marcinkevičius, V., and Dzemyda, G. (2017). A new web-based solution for modelling data mining processes. *Simulation Modelling Practice and Theory*.

Mell, P. and Grance, T. (2011). The NIST definition of cloud computing. *NIST Special Publication*, 800:145.

Michalski, S. R., Carbonell, G. J., and Mitchell, M. T. (1983). Machine learning: An artificial intelligence approach. *Understanding the Nature of Learning*, 2:3–26.

Nalchigar, S. and Yu, E. (2017). Conceptual modeling for business analytics: A framework and potential benefits. In *Business Informatics (CBI), 2017 IEEE 19th Conference on*, volume 1, pages 369–378. IEEE.

Naous, D., Schwarz, J., and Legner, C. (2017). Analytics as a service: Cloud computing and the transformation of business analytics business models and ecosystems. In *Proceedings of the 25th European Conference on Information Systems (ECIS), Guimares, Portugal*, pages 487–501.

Narman, P., Johnson, P., Ekstedt, M., Chenine, M., and Konig, J. (2009). Enterprise architecture analysis for data accuracy assessments. In *Enterprise Distributed Object Computing Conference, 2009. EDOC'09. IEEE International*, pages 24–33. IEEE.

Pruengkarn, R., Wong, K. W., and Fung, C. C. (2016). Data cleaning using complementary fuzzy support vector machine technique. In *International Conference on Neural Information Processing*, pages 160–167. Springer.

Pyle, D. (1999). *Data preparation for data mining*, volume 1. Morgan Kaufmann.

Ribeiro, M., Grolinger, K., and Capretz, M. A. (2015). Mlaas: Machine learning as a service. In *Machine Learning and Applications (ICMLA), 2015 IEEE 14th International Conference on*, pages 896–902. IEEE.

Riedemann, C. and Timm, C. (2003). Services for data integration. *Data Science Journal*, 2:90–99.

Rong, S., Niu, X., Xiang, E. W., Wang, H., Yang, Q., and Yu, Y. (2012). A machine learning approach for instance matching based on similarity metrics. In *International Semantic Web Conference*, pages 460–475. Springer.

Seibold, M. and Kemper, A. (2012). Database as a service. *Datenbank-Spektrum*, 12(1):59–62.

Shearer, C. (2000). The crisp-dm model: the new blueprint for data mining. *Journal of data warehousing*, 5(4):13–22.

Song, I.-Y., Zhu, Y., Ceong, H., and Thonggoom, O. (2015). Methodologies for semi-automated conceptual data modeling from requirements. In *International Conference on Conceptual Modeling*, pages 18–31. Springer.

Sun, Z., Strang, K., and Firmin, S. (2017). Business analytics-based enterprise information systems. *Journal of Computer Information Systems*, 57(2):169–178.

Terzo, O., Ruiu, P., Bucci, E., and Xhafa, F. (2013). Data as a service (daas) for sharing and processing of large data collections in the cloud. In *Complex, Intelligent, and Software Intensive Systems (CISIS), 2013 Seventh International Conference on*, pages 475–480. IEEE.

Vargo, S. L. and Lusch, R. F. (2004). Evolving to a new dominant logic for marketing. *Journal of marketing*, 68(1):1–17.

Vu, Q. H., Pham, T.-V., Truong, H.-L., Dustdar, S., and Asal, R. (2012). Demods: A description model for data-as-a-service. In *Advanced Information Networking and Applications (AINA), 2012 IEEE 26th International Conference on*, pages 605–612. IEEE.

Wagner, M. (2015). Integrating explicit knowledge in the visual analytics process. *Doctoral Consortium on Computer Vision, Imaging and Computer Graphics Theory and Applications (DCVISIGRAPP 2015), Scitepress Digital Library*.

Webster, J., Kingston, O., and Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly*, 26(2):xiii–xxiii.

Xinhua, E., Han, J., Wang, Y., and Liu, L. (2013). Big data-as-a-service: Definition and architecture. In *Communication Technology (ICCT), 2013 15th IEEE International Conference on*, pages 738–742. IEEE.

Xu, D., Wu, D., Xu, X., Zhu, L., and Bass, L. (2015). Making real time data analytics available as a service. In *Proceedings of the 11th International ACM SIGSOFT Conference on Quality of Software Architectures*, pages 73–82. ACM.

Yu, L., Wang, S., and Lai, K. K. (2006). An integrated data preparation scheme for neural network data analysis. *IEEE Transactions on Knowledge and Data Engineering*, 18(2):217–230.

Zorrilla, M. and García-Saiz, D. (2013). A service oriented architecture to provide data mining services for non-expert data miners. *Decision Support Systems*, 55(1):399–411.